MULTIMEDIA UNIVERSITY

STUDENT ID NO

TABLE NO: ____________

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2019/2020

TDS3551 DATA MANAGEMENT

(All sections/groups)

23 OCTOBER 2019
9.00 a. m. – 11.00 a. m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of ten (10) printed pages including the cover page and two extra blank pages.
2. There are four (4) questions in this paper.
3. Answer ALL QUESTIONS.
4. All questions carry equal marks (25 marks) and the distribution of the marks for each sub-question is given.
5. Please write all your answers in the spaces provided in this question paper.

Question	Mark
1	
2	
3	
4	
Total	

Question 1

a) What are the four (4) benefits of the Open Data Initiative? Briefly explain them in your answer

[8 marks]

b) Briefly explain the factors that distinguish between a data analyst, data scientist and a data engineer.

[6 marks]

c) Data analytics often say that data can be broken into four dimensions known as the four Vs. List down these Vs and give a brief explanation on how each of them affect how data is managed.

[8 marks]

d) Provide one (1) example of with an appropriate description for each of the following sources of data.

i. databases

ii. files

iii. IoT devices

[3 marks]

Question 2

a) Using a diagram, briefly explain the different layers of the Lambda architecture

[10 marks]

b) The data stored using Lambda architecture is said to exhibit three new properties not found in a conventional database. What are these properties and why are they important?

[9 marks]

c) State and explain three (3) of the *desired properties* of the Big Data architecture

[6 marks]

Question 3

a) Explain the roles of the two main "*nodes*" in a typical Hadoop setup.

[6 marks]

b) Hadoop 2.0 deployments encountered massive overhead costs in terms of storage. Explain this overhead and why it occurs.

[6 marks]

c) How does *erasure coding* in Hadoop 3.0 improve on the storage overhead costs?

[3 marks]

d) Explain the notion of block sizes as implemented in HDFS and why it contributes to the small file problem.

[6 marks]

e) What is the reason for having *heartbeats* in the Hadoop cluster? Explain what happens when there are problems with these *heartbeats*.

[4 marks]

Question 4

a) How does a relational database accommodate increases in data volume and velocity? Explain the approaches taken and the problems faced by doing so.

[6 marks]

b) During the process of data ingestion, data normalization is sometimes necessary. This normalization may possibly result in slower processing. Explain the normalization process and, using examples if necessary, why this slow down occurs.

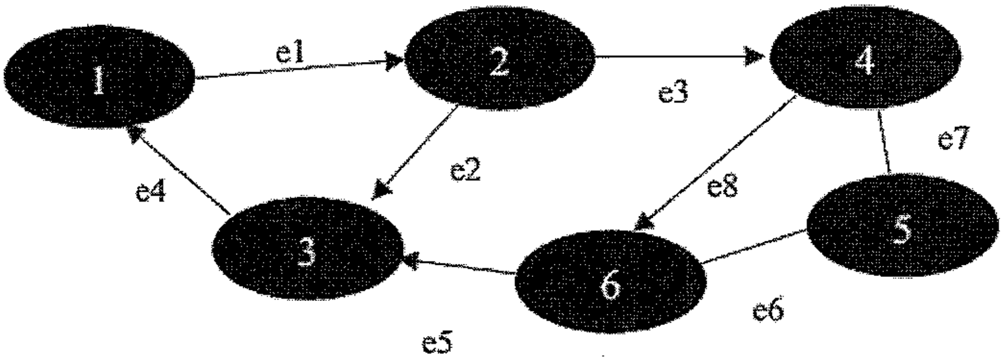
[4 marks]

c) Convert the data found in the following relational table into its equivalent fact-based model for the employee named "*Bob*".

Employee information						
ID	name	dateOf Birth	gender	wage	role	Timestamp
2333	Bob	10-10-80	M	1500	cleaner	30/7/2017
2565	Ina	10-04-83	F	1750	developer	3/6/2017
9982	Muthu	13-06-76	M	2000	secretary	15/10/2018
2565	Ina	10-04-83	F	1750	spy	31/9/2017
8844	Dave	16-11-76	M	3000	developer	3/7/2016
2333	Bob	10-10-80	M	1500	coffeeboy	31/8/2018
2333	Bob	10-10-80	M	2000	manager	1/5/2019
8844	Dave	16-11-76	M	3000	saboteur	30/9/2018

[8 marks]

d) Given the following graph below, write out the corresponding GraphML code declaration.



[7 marks]

End of paper